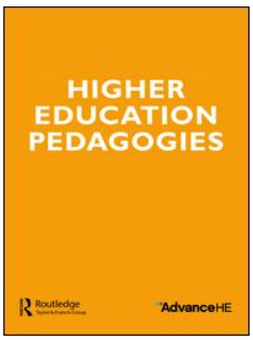
Threats to the validity of the Collegiate Learning Assessment (CLA+) as a measure of critical thinking skills and implications for Learning Gain



Higher Education Pedagogies

ISSN: (Print) 2375-2696 (Online) Journal homepage: http://www.tandfonline.com/loi/rhep20

Threats to the validity of the Collegiate Learning Assessment (CLA+) as a measure of critical thinking skills and implications for Learning Gain

Cesare Aloisi & A. Callaghan

To cite this article: Cesare Aloisi & A. Callaghan (2018) Threats to the validity of the Collegiate Learning Assessment (CLA+) as a measure of critical thinking skills and implications for Learning Gain, Higher Education Pedagogies, 3:1, 57-82, DOI: 10.1080/23752696.2018.1449128

To link to this article: https://doi.org/10.1080/23752696.2018.1449128

Published online: 06 Sep 2018.

HIGHER EDUCATION PEDAGOGIES
2018, VOL. 3, NO. 1, 57–82
https://doi.org/10.1080/23752696.2018.1449128

Threats to the validity of the Collegiate Learning Assessment (CLA+) as a measure of critical thinking skills and implications for Learning Gain

Cesare Aloisi‡ and A. Callaghan

University of Reading, Reading, UK

ABSTRACT

The University of Reading Learning Gain project is a three-year longitudinal project to test and evaluate a range of available methodologies and to draw conclusions on what might be the right combination of instruments for the measurement of Learning Gain in higher education. This paper analyses the validity of a measure of critical thinking skills, the Collegiate Learning Assessment (CLA+) and the implications of using this standardised test as a proxy for Learning Gain. The paper reviews five inferences regarding the interpretations and use of test scores: construct representation, scoring, generalisation, extrapolation and decision-making. Each section reviews some of the available evidence in support of the claims the CLA+ makes and the threats to their validity. The possible impact of these issues on Learning Gain in the UK is considered.

ARTICLE HISTORY
Received 2 October 2017
Revised 3 February 2018
Accepted 18 February 2018

KEYWORDS
Learning gain; collegiate learning assessment; critical thinking; validity; reliability

Introduction

The 2016 UK Government White Paper on the Teaching Excellence Framework (TEF) proposed that teaching and learning excellence will be measured by considering teaching quality, the learning environment, student outcomes (attainment) and learning gain. The latter is broadly defined by the Higher Education Funding Council for England (HEFCE) as 'an attempt to measure the improvement in knowledge, skills, work-readiness and personal development made by students during their time spent in higher education' (Higher Education Funding Council for England [HEFCE], 2016). A good learning gain measure should meet four key requirements: longitudinal or cross-sectional design; validity; representativeness; and comparability across disciplines, institutions and countries (McGrath, Guerin, Harte, Frearson, & Manville, 2015).

October 2015 saw the launch of a three-year, HEFCE-funded project on learning gain at the University of Reading. The project is one of 13 collaborative projects being launched over 70 universities in England. The specific aim of the University of Reading's three-year project is to test and evaluate a range of available methodologies (including grades, surveys

CONTACT Cesare Aloisi caloisi@aqa.org.uk
‡Centre for Education Research and Practice, Assessment and Qualifications Alliance.

and standardised tests) to draw conclusions on what might be the right combination of instruments for the measurement of learning gain in higher education. Outcomes of this project will feed into ongoing debates about the quality and impact of higher education, and how we evidence the value of investment in it.

One of the measures selected for trialling and evaluation was a standardised test of critical thinking skills. Critical thinking skills are thought to play a central role in logical thinking, decision-making and problem solving and any improvement in them following three years in a higher education institution (HEI) could be seen as a learning gain (Liu, Frankel, & Roohr, 2014). While it is debatable whether critical thinking is linked to any practical impact on academic attainment or career prospects, it has received considerable attention recently in the UK HE sector and has been included as one of the core learning outcomes by many HEIs.

The standardised test we chose to use was the *Collegiate Learning Assessment* (CLA), a proprietary test developed in 2002 by the Council for Aid to Education (CAE) to 'use real-world problem-solving tasks to measure students' critical-thinking skills' (Council for Aid to Education [CAE], 2015a). The CAE is a non-profit corporation established in 1952 in New York to increase private support to higher education with a view to increase student access. Between 1996 and 2005, it was a subsidiary of the RAND Corporation. The CLA was conceived to provide HEIs with a measure of student progression which went beyond academic skills and knowledge, and provided employers with a transferrable measure of work readiness (Benjamin et al., 2013). Because of this, it was deemed to be an appropriate tool to address the objectives of the learning gain project at Reading and the wider objectives of the HEFCE/Government agenda. That said, part of the research involved challenging these very objectives and the policy drive towards employability, whose assumptions can and have been called into question (Frankham, 2016; Winterbotham, Vivian, Shury, Davies, & Kik, 2014).

The specific version of the assessment trialled at Reading was introduced by the CAE in 2013/2014 and is called CLA+. The CLA+ is a 90-min (maximum) online assessment. It is composed of two parts: A Performance Task, which is a documentary analysis followed by an argumentative essay; and Selected-Response Questions (SRQ), a 30-min multiple-choice questionnaire. There are three subsections to this section; 'designed to measure [… the students'] ability to apply scientific and quantitative reasoning, critically read and evaluate the texts, and detect logical flaws and questionable assumptions to critique an argument' (CAE, 2014a).

The main difference between the original CLA and the CLA+ is that the focus of the former was 'the institution (rather than the student) as the unit of analysis' (Klein, Benjamin, Shavelson, & Bolus, 2007, p. 418), whereas the CLA+ is claimed to be sufficiently reliable to be used both at the institutional and at the student level (Zahner, 2014a). In practice, the CLA did not have a 'Selected-Response' section and, in each institution, different samples of students would be assigned one of eight 'Performance Tasks' so that all tasks would be administered on an institutional level. The emphasis of this article is on the validity of the CLA+ as a tool to measure Learning Gain.

Validity is the degree to which the proposed meaning of test outcomes and uses of the test are warranted by its qualities and justified within the context in which it is administered (Messick, 1989; see Newton & Shaw, 2015; for a review of the evolution of the concept). For example, a high score in the CLA+ is taken to mean that a student can think 'critically' and will be able to perform certain tasks in a range of settings. HEIs are told that the CLA+ will help them to detect deficits in work readiness and target future instruction (CAE, 2017).

These claims are supported by a network of assumptions, inferences and arguments regarding the ability of the test to measure certain skills accurately, objectively and consistently thanks to its technical properties and the way it is administered and marked.

A comprehensive validation of the CLA+ would require making this inferential network evident, analysing the evidence in its support and evaluating which assumptions stand to scrutiny and which do not. This is too broad a scope for a single research article; it would require a wide and deep analysis drawing from multiple sources of evidence that are sometimes unavailable. One example of full validation is the validation of the A-level Physics qualification by Cambridge Assessment (Shaw & Crisp, 2012), which included a range of internal and tailored evidence such as item-level analyses, assessment policy reviews and expert panels.

This article adopts Shaw and Crisp's (2012) approach to validation but is more modest in scope, analysing only a selection of threats to the validity of the CLA+. In the spirit of fairness, only features of the assessment that these authors think could be readily reviewed by the CAE are reported. The framework is based on Kane's (2013) understanding of validation, whereby what is validated are arguments concerning interpretations and uses of test scores.

The main claim about the interpretation of CLA+ scores is that they represent skill levels in critical thinking and written communication. This is equivalent to claiming that a candidate's unobservable critical thinking skills can be inferred by his or her unique attempt. The validity of this claim is founded on an inferential chain comprising the following steps, or *inferences* (Kane, 2013; Shaw & Crisp, 2012):

(1) *Construct Representation*: performance on the test implies performance on the construct (see the relevant section for a definition of 'construct').
(2) *Scoring*: score differences capture performance differences.
(3) *Generalisation*: one set of scores can serve as a general estimate of expected performance for any equivalent version of the test.
(4) *Extrapolation*: the competence in the construct expressed by the test performance can be applied to larger domains and new situations.
(5) *Decision-making*: different levels of competence can or should lead to different decisions about the candidates.

Each inference relies on a *warrant*, 'a statement that is claimed to be true and justifies the related inference if appropriately supported by evidence' (Shaw & Crisp, 2012, p. 8). Taking the Construct Representation inference as an example, one can state that performance on the test implies performance on the construct only if the test actually elicits performances that build on the intended construct (warrant). The warrant should be based on evidence and depends on some assumptions (prerequisites for it to be true) such as the possibility to define and assess the construct.

Each section in this article reviews evidence in support of one of the five inferences above and seeks to identify gaps or contradictions that might threaten the validity of the CLA+. This is followed by a discussion on the suitability of the CLA+ as an instrument to detect learning gain in the current educational context.

This research is situated among a small number of other studies on the CLA/CLA+. Most of them are about assessment qualities (e.g. Klein, Liu, & Sconing, 2009; Zahner & Steedle, 2015) and are reviewed below. Another study (Steedle, 2014) looks at CLA outcomes to

explore new methods to apply motivation filtering (a procedure to dampen the effect of a minority of low-motivated students on average test scores) to standardised tests.

The most prominent publication to use the CLA as its primary source of evidence was a book by Arum and Roksa (Arum & Roksa, 2011), whose findings started an engaged debate in the media (Glenn, 2011) and within academia (Arum, 2013; Lindsay, 2013; Menchaca, 2014). The book was based on a report (Arum, Roksa, & Cho, 2011), showing an association between CLA scores and academic rigour in higher education (spending more hours studying alone, taking more challenging classes and being in a more demanding department). These findings are discussed in the 'Extrapolation' section.

Construct representation

The first element analysed in this article, and a pivotal concept in validity theory, is the *construct* underpinning the CLA+. A psychological construct is 'some postulated attribute of people, assumed to be reflected in test performance' (Cronbach & Meehl, 1955, p. 283). A construct is *not* what the scoring rubric awards marks for – i.e. it is not the type and level of performance that gets credited. It is a postulated but unobservable underlying ability causing the performance to take place.

The CLA/CLA+ are founded on the postulate that critical thinking (CT) skills can be defined and assessed. The evidence in support of these assumptions is reviewed below.

Construct definition

Supporting evidence

According to the CAE, CT are 'broad' and transferrable skills, but they are neither 'general reasoning abilities generally thought of as intelligence or *G*, nor […] the domain-specific skills limited to one or a few disciplines' (Benjamin et al., 2013, p. 6). The origin of this view can be traced to a framework for cognitive outcomes devised by Shavelson and Huang (2003) as an attempt to guide assessment design in a regime of high-stakes accountability. The framework did not mention CT skills directly, but it did introduce the concept of *broad abilities*: 'particular complexes of cognitive processes ('thinking') that underlie what we generally call verbal, quantitative, and spatial reasoning – as well as comprehension, problem-solving, and decision-making skills within [… and across] domains' (Shavelson & Huang, 2003, p. 15). This conceptualisation echoed Messick's 'broad cognitive abilities of comprehension, memory, visualisation, restructuring, reasoning, fluency' (Messick, 1984, p. 221).

Later research linked Shavelson and Huang's (2003) framework to extant CT tests (Klein, Kuh, Chun, Hamilton, & Shavelson, 2005), but little remains of this historical and theoretical heritage in recent CAE documents; today it is simply claimed that the CLA is 'well aligned' (p. 7) with three definitions of CT (Table 1).

Facione's (1990) definition is one of the most often quoted by this kind of assessments. It was the outcome of an enquiry that took two years, 46 experts (96% male) and employed the Delphi method (Dalkey & Helmer, 1962). The study was requested with a view of introducing and assessing a CT curriculum in the United States, covering pre-primary up to secondary education.

Table 1. The three definitions of critical thinking referred to by the CAE.

Facione (1990, p. 2)	Bok (2006, p. 109)	Pascarella and Terenzini (2005, p. 156)
We understand critical thinking to be purposeful, self-regulatory judgement which results in interpretation, analysis, evaluation and inference, as well as explanation of the evidential, conceptual, methodological, criteriological or contextual considerations upon which that judgement is based	The ability to think critically – ask pertinent questions, recognise and define problems, identify arguments on all sides of an issue, search for and use relevant data and arrive in the end at carefully reasoned judgments – is the indispensable means of making effective use of information and knowledge	Most attempts to define and measure critical thinking operationally focus on an individual's capability to do some or all of the following: identify central issues and assumptions in an argument, recognise important relationships, make correct references from the data, deduce conclusions from information or data provided, interpret whether conclusions are warranted based on given data, evaluate evidence of authority, make self-corrections and solve problems

The expert consensus built on the work of other philosophers and owed to Ennis (1962) article 'A concept of critical thinking', which is sometimes credited to have rekindled an interest on the topic (Thayer-Bacon, 1998). Both then and now, definitions of critical thinking abounded, with each author taking a comparable yet different stance on the matter. Table 2 summarises some of the definitions that might have informed Facione (1990), but many others are available (e.g. Elder, 2007; Halpern, 2013; see Lai, 2011, for a review).

There is substantial agreement in philosophy and psychology that CT involves both skills and the disposition to apply them (Lai, 2011), though the two need not co-occur. A thinking disposition indicates 'broad tendencies of pragmatic and epistemic self-regulation at a high level of cognitive control' (Evans & Stanovich, 2013, p. 230) and is a function of age, personality, cultural environment or formal education (Alexander, 2014; Murphy, Rowe, Ramani, & Silverman, 2014).

Critical thinking dispositions correlate with need for cognition (Stedman, Irani, Friedel, Rhoades, & Ricketts, 2009), which is 'a stable individual difference in people's tendency to engage in and enjoy effortful cognitive activity' (Cacioppo, Petty, Feinstein, & Jarvis, 1996, p. 198). In turn, need for cognition is related to metacognition (Coutinho, 2006) and often with some personality traits such openness, conscientiousness and (negatively) with neuroticism (Furnham & Thorne, 2013). The fact that both CT skills and CT disposition are correlated with known constructs in psychology made some authors question whether 'critical thinking' should be considered a stand-alone construct at all, rather than the outcome from the interaction of more established constructs (Stanovich, 2016).

Table 2. Various definitions of critical thinking in the rationalist philosophical tradition.

Definition	Reference
'The propensity and skill to engage in an activity with reflective skepticism'	McPeck (1981, p. 8)
'(1) it is *self-corrective* thinking; (2) it is thinking *with criteria*; and (3) it is thinking that is *sensitive to context*'	Lipman (1987, p. 5, emphasis in the text)
'Thinking that is appropriately moved by reasons'	Siegel (1988, p. 23)
CT 'is based on universal intellectual values that transcend subject matter divisions: clarity, accuracy, precision, consistency, relevance, sound evidence, good reasons, depth, breadth, and fairness'	Scriven & Paul (1987)
'Disciplined, self-directed thinking that exemplifies the perfections of thinking appropriate to a particular mode or domain of thought'	Paul (1992, p. 9)
'Reasonable, reflective thinking focused on deciding what to believe or do'	Ennis (1987, p. 10)

Threats to validity

These authors would argue, along with Johnson and Hamby (2015), that current definitions of CT are mutually exclusive, even though they all aim to situate themselves within the same theoretical tradition. Is CT about using strategies 'that increase the probability of a desirable outcome' (Halpern, 2013, p. 4), or is it about living 'rationally, reasonably, empathically' because of an awareness of the 'inherently flawed nature of human thinking when left unchecked' (Elder, 2007)?

The fact that the CLA+ is supposedly 'well aligned' (Benjamin et al., 2013, p. 7) with the definitions in Table 1 does not explain with sufficient clarity the *kind* of critical thinking the assessment tries to measure. Besides, even the definitions in Table 1 are at times at odds with each other:

- Bok and Pascarella and Terenzini (P&T) mention problem-solving, but Facione does not. Problem-solving is different from critical thinking (Bassok & Novick, 2012; Byrnes & Dunbar, 2014).
- Bok speaks about identifying all sides of an issue, P&T implicitly agree by mentioning 'assumptions', but this aspect is absent in Facione.
- P&T and Facione identify self-correction and self-regulation as a component of critical thinking, whereas this is left implicit in Bok.
- P&T raise the issue of credibility ('evidence of authority'), which is a 'criteriological consideration' in Facione but not an issue for Bok.
- Bok states that judgements should be 'carefully' reasoned, which is in line with the idea that CT should be 'effortful, […] mentally taxing' (Byrnes & Dunbar, 2014, p. 481), but this criterion is absent in the other two definitions.

Working from an explicit construct is crucial for test design. Having drawn from established tests, the precursors to the CLA (such as the performance tasks in Klein et al., 2005), might have done so. However, it is less clear if this is still the case today. In all reviewed documents, a CT construct is mentioned but the supporting information tends to concern only the qualities student work should exhibit. This is what gets credited in the mark scheme but, as already mentioned, it is not the construct.

Contrast this with Cambridge Assessment's approach. After many years of developing CT tests, Cambridge Assessment acknowledged that the many competing definitions of CT skills were affecting the clarity of the construct:

> It is perhaps fair to say that, in the absence of a single agreed definition in the area, the conception of what these tests measured had been largely transmitted implicitly through the coincidence of a core group of common experts and personnel working on these tests and writing items for them. (Black, 2012, p. 124)

Therefore, Cambridge Assessment carried out research 'to create a definition and taxonomy of Critical Thinking in order to support validity arguments about Critical Thinking tests and exams' (Black, 2012, p. 124). Currently, a unique CT definition is available in research articles (see Black, 2012, p. 125) but also to the wider public (see Cambridge International Examinations, 2016). Why is not there a shared and explicit definition of CT across CAE's documents?

The CLA+ also purports to capture effective writing, but again without defining the construct explicitly or articulating the relationship between effective writing and CT. In older documents, they were presented as separate but complementary educational outcomes: 'the

CLA was designed to test a student's critical thinking, analytic reasoning, problem solving, *and* written communications competencies' (Klein et al., 2007, p. 417, emphasis added). In recent documents, critical thinking seems to have subsumed the other skills: 'more emphasis is placed on critical-thinking skills, *such as* analytic and quantitative reasoning, problem-solving, and written communication' (CAE, 2014b, p. 1, emphasis added).

Clark and Watson (1995) have argued that without 'an articulated theory […] there is no construct validity'; therefore, a 'critical first step is to develop a precise and detailed conception of the target construct and its theoretical context' (p. 310). In the case of the CLA+, the conceptions of CT or effective writing are neither precise nor detailed. Instead, the CAE seems to assume that there is widespread agreement of what CT is, that the test measures it and that test outcomes are an accurate quantification of the ability. From there, the CAE proceeds by laying out expected responses and marking criteria. The internal logic is valid: if there is a one-to-one correspondence between the construct and what the test measures, then stating what the test measures is a sufficient descriptor of the construct. In other words, the implicit position of the CAE is that critical thinking is what the CLA+ tests.

This approach towards psychological measurement was common in the 1910–1920s (Kane, 2013; Sireci, 1998), but it is inadequate today and does not reflect the lack of consensus regarding the theorisation of CT. Even if the internal logic in CAE's argument is valid, it may not be sound[1]: it is not known whether there is a one-to-one correspondence between the construct and what the test measures, this is matter of validation. The position whereby critical thinking is what the CLA+ tests is untenable, as it assumes a priori what should be deduced a posteriori.

Assessment of the construct

Supporting evidence

There are currently several standardised CT tests (Table 3).

The CLA (not the CLA+) was found to correlate with some of these tests, which is traditionally one of the pieces of evidence used to infer that two tests are measuring the same construct (it used to be called 'concurrent validity'; Cronbach & Meehl, 1955). Specifically, the *Test Validity Study* (Klein et al., 2009) compared the CLA with the MAPP (now known as the Proficiency Profile), which was Educational Testing Service's Measure of Academic Proficiency and Progress (ETS, 2017); and with the CAAP, ACT's Collegiate Assessment of Academic Proficiency (ACT, 2017).

After sampling over 500 students from 13 institutions, the correlations between the critical thinking components of the MAPP and the CAAP, and the CLA, were 0.53–0.58 at the student level and 0.79–0.83 at the institutional level.

Threats to validity

The first point to note about the assessments in Table 3 is that, while they all claim to be measuring CT skills or dispositions, they organise them into different taxonomies. This is a consequence of the fragmented theoretical landscape mentioned earlier: having several but different CT assessments may be evidence that the construct can be captured from different angles, or on the contrary that different constructs are being captured. For example, an attempt to apply CLA+ principles in a comparative setting found that 'the Australian Council of Educational Research's (ACER) analytic-reasoning test assesses different abilities

Table 3. Other instruments purporting to measure 'critical thinking' skills.

Instrument	Format	Time to complete	Subscales	Reference paper
Watson–Glaser Critical Thinking Appraisal	80 MC items (long version); 41 MC items (short version); Forms A & B (pre-post-test)	30–40/45–60	Inference; Recognition of assumptions; Deduction; Interpretation; Evaluation of arguments	Watson and Glaser (1980)
California Critical Thinking Skills Test	34-item multiple choice; Forms A & B (pre- post-test)	45 min	Analysis; Evaluation; Inference; Inductive reasoning; Deductive reasoning	Facione (1990)
Ennis–Weir Critical Thinking Essay Test	Essay	40 min	Getting the point; Seeing reasons and assumptions; Stating one's point; Offering good reasons; Seeing other possibilities; Responding appropriately and/or avoiding legal arguments	Ennis and Weir (1985)
Cornell Critical Thinking Test	Level X (Grade 5–12+) = 71 MC; Level Z (Grades 11–12+) = 52 MC	50 min	Induction; Deduction; Value judgement; Observation; Credibility; Assumptions; Meaning + semantics, definition and prediction in planning experiments	Ennis, Millman, and Tomko (1985)
California Critical Thinking Disposition Inventory	75 items, six-point Likert	15–20 min	Open-mindedness; Self-confidence; Maturity; Analyticity; Systematicity; Inquiry; Truth seeking	Facione and Facione (1992)
Performance-based Development System	Constructed response	240	Critical thinking; Interpersonal skills; technical skills	Del Bueno (1990)
Critical Thinking Diagnostic	25 items, six point Likert	15	Problem recognition; Clinical decision-making; Prioritisation; Clinical application; Reflection	Berkow, Virkstis, Stewart, Aronson, and Donohue (2011)
Thinking Skills Assessment	50 MC	90	Problem-solving skills, including numerical and spatial reasoning. Critical thinking skills, including understanding argument and reasoning using everyday language.	Cambridge Assessment Admissions Testing (2017)
Halpern Critical Thinking Assessment	25 items, both CR and MC (or MC only, short version)	60–80/20	Verbal reasoning; argument analysis; thinking as hypothesis testing; likelihood and uncertainty; decision-making and problem solving	Halpern (2010)
California Measure of Mental Motivation (CM3)	72 items, 4-pt Likert	20	(a) learning orientation, (b) creative problem solving, (c) cognitive integrity, (d) scholarly rigor and (e) technological orientation	Insight Assessment (2013)
Collegiate Assessment of Academic Proficiency	32 MC	40	Clarifying, analysing, evaluating and extending arguments	ACT (2017)

(*Continued*)

Table 3. (*Continued*).

Instrument	Format	Time to complete	Subscales	Reference paper
ETS Proficiency Profile (EPP) Critical Thinking	27 MC	40	(a) distinguish between rhetoric and argumentation in a piece of nonfiction prose, (b) recognise assumptions and the best hypothesis to account for information presented, (c) infer and interpret a relationship between variables and (d) draw valid conclusions based on information presented	ETS (2017)
Discussion board analysis	NA	60	Analysis, inference, interpretation, explanation, evaluation and self-regulation	Pucer et al (2014)

Notes: There are other CT tools that are specific to nursing but it is not clear whether they measure domain-specific constructs (e.g. Carter et al., 2015; Romeo, 2010). Likewise, some instruments were developed to measure critical-analytic skills (Lawson, Jordan-Fleming, & Bodle, 2015). Cambridge Assessment has a range of other instruments, including GCSE/A-level critical thinking tests (Black, 2012). New instruments are being developed. Examples include: the Critical Thinking Toolkit (Stupple et al., 2017); the Critical Thinking Disposition Scale (Sosu, 2013); or ETS's HEIghten™ (Liu, Mao, Frankel, & Xu, 2016).

Sources: Brunt (2005), Carter, Creedy, and Sidebotham (2015), Liu et al. (2014), O'Hare and McGuinness (2015), and Zuriguel Pérez et al. (2015).

than CAE's [… SRQ] even though both tests are billed as measuring the 'same' construct' (Wolf, Zahner, & Benjamin, 2015, p. 473).

Regarding the *Test Validity Study*, CAE's researchers highlighted the positive correlations between the CLA and 'other tasks that measure critical thinking' (Zahner, 2014a, p. 4), but they did not mention that high correlations were observed across a range of domains. At the student level, the CLA correlated with measures of critical thinking as strongly as it did with measures of reading skills, science and, to a lesser degree, writing and mathematics skills (Klein et al., 2009, Tables 2a and 2b, p. 24). At the institutional level, critical thinking correlations were as strong as correlations with science and some measures of writing, and even *higher* correlations were recorded with reading and mathematics (0.76–0.91).

Similarly, a large sample of over 10,000 first-year students from 113 institutions in 2005 and over 4000 finalists from 90 institutions in 2006 showed that the CLA had student-level correlations in the range of 0.54–0.56 and institution-level correlations of 0.88–0.91 with the SAT, a mathematics and home language – not CT – test used in the US for university admission purposes (College Board, 2017).

Even Arum and Roksa (2011) interpreted the greater CLA gains of students graduating in liberal arts compared to students in business, education, social work and communication as being due to higher quality and greater quantity of reading and writing, rather than better CT skills (notice, however, that after two years the correlation between CLA and programme of study appeared to be substantively moderated by student socio-economic background and institution; Arum & Roksa, 2011, Table A4.3).

These findings invite once again the question about what exactly the CLA/CLA+ measures. How is 'critical thinking' different from the general academic ability underpinning reading, mathematics and science literacy?

Scoring

The *scoring inference* concerns the rules quantifying a candidate's observed performance. Assumptions include the scoring system (policies, rules, marking criteria) being fit-for-purpose, or that variations in student performance should depend on the construct and not on other confounding factors. For the sake of conciseness, only the first point will be covered in this analysis, but evidence of the second point was also produced.

Adequacy of the scoring system

Supporting evidence

The first part of the test is the performance task (PT), an argumentative essay informed by a documentary analysis.

The scoring rubric is divided into three subscales (often referred to as 'marking criteria' in the literature; see Popham, 1997), each situating one aspect of student performance on one of six increasing proficiency levels carrying 1–6 marks (a mark of 0 flags the test for exclusion and the student does not receive a PT score). The subscales are: Analysis and Problem Solving, Writing Effectiveness and Writing Mechanics. Each level on each scale is associated to a performance criterion (Table 4). A marker's task is to determine which criteria are met and to award marks accordingly.

Each PT is double-marked on each subscale, the two marks are averaged and then added across the three subscales. Klein et al. (2007) explain that the PT was modelled after the California bar examination (Klein, 1996) and the Tasks in Critical Thinking (Erwin & Sebrell, 2003), but the initial scoring rubric was different than the current one, comprising 40 dichotomous items and a 5-point communication score (Klein, 2008). It was not possible to identify when and how the current scoring rubric was developed.

The second part of the test is the SRQ, and it is further divided into three subsections carrying 10, 10 and 5 marks. Once the raw PT and the SRQ marks are scaled, the total CLA+ score is the average of the scaled PT and SRQ scores. A student's mastery level (Below basic, Basic, Proficient, Accomplished and Advanced) depends on this score. Table 5 reproduces the Proficient level descriptor for reference.

Threats to validity

A first issue with the scoring system lies in adding marks from different subscales. Klein et al. (2007) argued that, unlike other competing assessments, the CLA+ PT 'recognizes that critical thinking, analytic reasoning, problem solving, and written communication skills are inherently and complexly intertwined in the task and response demands' (p. 421). However, the analytic approach of the scoring rubric, where performance is analysed along three separate dimensions that are then added together, is precisely the 'stitching' of components that Klein et al. (2007) claim to have avoided. They state that 'the whole is usually much greater than the sum of its parts' (p. 422), yet to score the PT one just has to sum the parts to get the whole. This additive rule does not seem to have any 'sound theoretical rationale', and it is 'essentially a device of convenience' (Sadler, 2009, p. 171).

A second issue is with the practical consequences of such scoring system. By allowing for extreme opposite performances on different criteria to average out (Sadler, 2009), marks will naturally converge towards the mean. Since there are 271 ways to receive a total mark between 10 and 11, but only two to be awarded 3 or 18, there will be a statistical tendency for the mid-range marks to attract most results.

This probability is affected by other factors. For instance, level descriptors have some skill overlap, in the sense that a good mark in one criterion tends to call for a similar mark in the others. There are also more ways to capture an unsatisfactory performance than a good one, because the level meant to represent a satisfactory performance is level 4 out of 6. Even considering these factors, however, central values would appear more often (there are more numbers that could reasonably follow two 4s than two 1s).

Regarding the SRQs, they were introduced around 2012 'to improve the precision of student-level results' (Zahner, 2014a, p. 1). This reinforces the critique made above that there is no specific theoretical reason for adding and averaging scores. The SRQ was certainly developed with a view of capturing similar skills to the PT, but the purpose of the section was to increase reliability. There is no guarantee that the test has become more valid.

Finally, when it comes to the total score, it can be shown that the linear transformations used to scale raw scores reward good writing over CT, in the sense that a student delivering a good performance in the PT (15–15.5 points) and *randomly guessing* all questions in the SRQ would stand a good chance of being considered Proficient.

For example, imagine a student with a solid command of written English, able to cite a few sources and expand on the answer, but who is also somewhat biased and does not understand some finer details. Following the rubric in Table 4, a fair mark could be: 4 + 5 + 6 = 15.

Table 4. CLA+ mark scheme.

	1	2	3	4	5	6
Analysis and Problem Solving	May state or imply a decision/conclusion/ position	States or implies a decision/conclusion/ position	States or implies a decision/conclusion/ position	States an explicit decision/conclusion/ position	States an explicit decision/conclusion/ position	States an explicit decision/conclusion/ position
Making a logical decision or conclusion (or taking a position) and supporting it by utilising appropriate information (facts, ideas, computed values or salient features) from the Document Library	Provides minimal analysis as support (e.g. briefly addresses only one idea from one document) or analysis is entirely inaccurate, illogical, unreliable or unconnected to the decision/conclusion/ position	Provides analysis that addresses a few ideas as support, some of which is inaccurate, illogical, unreliable or unconnected to the decision/conclusion/ position	Provides some valid support, but omits or misrepresents critical information, suggesting only superficial analysis and partial comprehension of the documents	Provides valid support that addresses multiple pieces of relevant and credible information in a manner that demonstrates adequate analysis and comprehension of the documents; some information is omitted	Provides strong support that addresses much of the relevant and credible information, in a manner that demonstrates very good analysis and comprehension of the documents	Provides comprehensive support, including nearly all of the relevant and credible information, in a manner that demonstrates outstanding analysis and comprehension of the documents
			May not account for contradictory information (if applicable)	May attempt to address contradictory information or alternative decisions/conclusions/ positions (if applicable)	Refutes contradictory information or alternative decisions/ conclusions/positions (if applicable)	Thoroughly refutes contradictory evidence or alternative decisions/ conclusions/positions (if applicable)
Writing Effectiveness	Does not develop convincing arguments; writing may be disorganised and confusing	Provides limited, invalid, over-stated, or very unclear arguments; may present information in a disorganised fashion or undermine own points	Provides limited or somewhat unclear arguments. Presents relevant information in each response, but that information is not woven into arguments	Organises response in a way that makes the writer's arguments and logic of those arguments apparent but not obvious	Organises response in a logically cohesive way that makes it fairly easy to follow the writer's arguments	Organises response in a logically cohesive way that makes it very easy to follow the writer's arguments
Constructing organised and logically cohesive arguments. Strengthening the writer's position by providing elaboration on facts or ideas (e.g. explaining how evidence bears on the problem, providing examples, and emphasising especially convincing evidence)	Does not provide elaboration on facts or ideas	Any elaboration on facts or ideas tends to be vague, irrelevant, inaccurate or unreliable (e.g. based entirely on writer's opinion); sources of information are often unclear	Provides elaboration on facts or ideas a few times, some of which is valid; sources of information are sometimes unclear	Provides valid elaboration on facts or ideas several times and cites sources of information	Provides valid elaboration on facts or ideas related to each argument and cites sources of information	Provides valid and comprehensive elaboration on facts or ideas related to each argument and clearly cites sources of information

(*Continued*)

Table 4. (*Continued*).

	1	2	3	4	5	6
Writing Mechanics	Demonstrates minimal control of grammatical conventions with many errors that make the response difficult to read or provides insufficient evidence to judge	Demonstrates poor control of grammatical conventions with frequent minor errors and some severe errors	Demonstrates fair control of grammatical conventions with frequent minor errors	Demonstrates good control of grammatical conventions with few errors	Demonstrates very good control of grammatical conventions	Demonstrates outstanding control of grammatical conventions
Demonstrating facility with the conventions of standard written English (agreement, tense, capitalisation, punctuation and spelling) and control of the English language, including syntax (sentence structure) and diction (word choice and usage)	Writes sentences that are repetitive or incomplete and some are difficult to understand Uses simple vocabulary and some vocabulary is used inaccurately or in a way that makes meaning unclear	Consistently writes sentences with similar structure and length and some may be difficult to understand Uses simple vocabulary and some vocabulary may be used inaccurately or in a way that makes meaning unclear	Writes sentences that read naturally but tend to have similar structure and length Uses vocabulary that communicates ideas adequately but lacks variety	Writes well-constructed sentences with some varied structure and length Uses vocabulary that clearly communicates ideas but lacks variety	Consistently writes well-constructed sentences with varied structure and length Uses varied and sometimes advanced vocabulary that effectively communicates ideas	Consistently writes well-constructed complex sentences with varied structure and length Displays adept use of vocabulary that is precise, advanced and varied

Source: Reproduced verbatim from CAE (2015b).

Table 5. The proficient mastery level descriptor.

Students at the proficient level should be able to extract the major relevant pieces of evidence provided in the documents and provide a cohesive argument and analysis of the task. Proficient students should be able to distinguish the quality of the evidence in these documents and express the appropriate level of conviction in their conclusion given the provided evidence. Additionally, students should be able to suggest additional research and/or consider the counterarguments. Minor errors in writing need to be defined rigorously
Proficient students have the ability to correctly identify logical fallacies, accurately interpret quantitative evidence, and distinguish the validity of evidence and its purpose. They should have the ability to determine the truth and validity of an argument. Finally, students should be able to know when a graph or table is applicable to an argument

Source: Reproduced verbatim from CAE (2015b).

If this student randomly guesses all questions throughout the SRQ, he or she will have a 27% chance of getting *at least* 3, 2 and 1 correct answers per section and have a total CLA+ score of about 1084 points. This is very close to the low Proficient boundary (the exact cut scores are not known, but see the graphs in CAE, 2015b, p. 3), and it takes very little (a half-point more in the PT, one extra point in the SRQ) to cross the threshold.

A student scoring 17 in the PT only needs to get two correct answers in the first SRQ section and one in the second (probability by random guessing, 71%), then he or she can skip the third section completely and still accumulate enough points to qualify for Proficient. Any such student would not have demonstrated to meet any of the criteria in the second paragraph of the mastery level descriptor (Table 5). Yet, an employer could look at the mastery level (a level of Proficient or higher allows students to receive a digital badge for their curriculum vitae) and be led to think the student is able to identify logical fallacies, take different viewpoints and interpret quantitative evidence accurately.

To get a sense of what this means in practice, 8300 finalists took part in the CLA+ in the US in 2015/2016 (CAE, 2016, Table 3). Assuming a joint probability distribution in PT marks in absence of actual data, about 461 would have received a mark between 15 and 16.5, and 62 a mark at or above 17. Had they all decided to answer the SRQ randomly, this would have resulted in over 200 students (2.5% of all finalists) being labelled Proficient whilst having completed in practice only half of the test.

There were 480,575 graduates in the UK in 2015/2016 (HEFCE, 2017). In this hypothetical scenario, over 12,000 would have been labelled Proficient despite guessing, and a much larger percentage of students could achieve a Basic level *in critical thinking and problem solving* by relying on good writing skills and citing a few documents.

Of course, these numbers are speculative. The actual point being made is that scores do not translate well into performance levels; the inferences may be unwarranted. When writing accounts for a third of the total marks and the ability to write a convincing argument has a greater weight than being able to show *why* certain evidence is more credible, the total score can easily stop having its proposed meaning.

Generalisation

From a test score, one generally infers that a candidate would perform similarly if administered a comparable version, that is, the score should be representative of a candidate's ability. This relies on the assumption that the test is reliable across administrations and over time.

Table 6. Reliability information on the CLA and CLA+.

Type	Correlation	Contextual information
Split-sample correlation		
	0.94	Instrument: CLA
	(first-year)	*Outcome*: Mean scores
	0.86	*Sample*: 62 and 44 institutions, 40 students per sample minimum, 2005–2006
	(finalists)	*Source*: Klein et al. (2007)
	0.85	Instrument: CLA
	(first-year)	*Outcome*: Mean scores
	0.64	*Sample*: 13 institutions, fewer than 30 students per sample, 2008
	(finalists)	*Source*: Klein et al. (2009, p. 29)
	0.77	Instrument: CLA
	(first-year)	*Outcome*: Residuals
	0.70	*Source*: Klein et al. (2007)
	(finalists)	
	0.74	Instrument: CLA
	(pooled, 2008)	*Outcome*: Residuals
	0.75	*Sample*: 150 and 140 institutions, 25 students per sample minimum, 2007–2009
	(pooled, 2009)	*Source*: Steedle (2012, p. 644)
Year-to-year consistency		
	0.32–0.55	Instrument: CLA *Outcome*: Residuals *Sample*: 87 institutions participating in the 2007/2008 and 2008/2009 cycles *Source*: Steedle (2012, p. 645)
	0.51–0.53	Instrument: CLA+ *Outcome*: Residuals *Sample*: 25 institutions, cross-sectional data collected in 2005/2006 and longitudinal data collected between 2005 and 2009 *Source*: Zahner and Steedle (2015, p. 7)

Test reliability

Supporting evidence

The CLA+ has high values of Cronbach's alpha (0.81; CAE, 2014a, p. 5; 0.85–0.87; Zahner, 2014a, p. 2). This is a measure of between-item correlation; high values suggest that 'there is little variance specific to individual items' (Cortina, 1993, p. 100).

Moderate-to-strong inter-rater correlations, summarising the extent to which the set of scores assigned by two markers agree, provide evidence in support of the claim that it is possible to ensure consistent scoring across markers. The correlations range from 0.67–0.75 (Zahner, 2014a, p. 2) to 0.80–0.88 (CAE, 2014a, p. 5; Klein et al., 2007, p. 429). These values are in line with some public examinations in England (Opposs & He, 2011).

Reliability evidence also comes by split-sample correlation studies: the sample is split into two subsamples, the mean of each is taken and then the means are correlated across institutions. Table 6 reports these correlations for both means and regression residuals using different methodologies, though note that they all refer to the older CLA, not the new CLA+.

Finally, it was possible to find evidence of year-to-year consistency (Table 6). This gives an indication of score reliability over time. One way to assess this would be to administer the same test more than once. There is a risk that follow-up scores could be inflated by test familiarity, but in fact this effect might disappear after one or two years (McKelvie, 1992), especially in the case of the SRQ. However, the CAE never trialled this type of test-retest

approach, and the information in the table is longitudinal only in the sense that the same university or student participated in two different test administrations.

Using the CLA, Steedle (2012) found that the residual scores of the same 87 institutions between to two consecutive years had a correlation of 0.55 (0.32, removing outliers). This is in line with the 0.51 coefficient found by Zahner and Steedle (2015) when they correlated cross-sectional and longitudinal residual scores using two comparable models.[2] These data suggest that university performance is very sensitive to the student sample: one university could find itself below statistical expectation one year and above the next year. Nevertheless, some universities produced very similar improvements in two student samples (see Fig. 1 in Zahner & Steedle, 2015).

Threats to validity

Cronbach's alpha should only be taken as a basic but insufficient requirement in modern testing (Barbaranelli, Lee, Vellone, & Riegel, 2015; Tavakol & Dennick, 2011). It does not mean that all items are testing the same construct (Cortina, 1993; Tavakol & Dennick, 2011). In this case, the values reported for the CLA+ are mostly driven up by the addition of the SRQ, which tends to reduce variability.

Inter-rater and split-sample mean-score correlations are adequate, but the residual correlations less so. This point was acknowledged by Steedle (2012), but it should be added that correlating level-2 residuals might not be statistically sound. They are not parameter estimates like a mean is, and there are still many unknowns in the literature about their distributional properties (Bates, 2007, 2010; Goldstein, 2011).

With regard to the year-to-year consistency, the CAE claims that all versions of the test are equivalent, but it was not possible to retrieve any supporting evidence. The strength of the correlation is only moderate and, while it is certainly possible that correlations captured differences in the CT skills of successive cohorts, changes in student recruitment protocols, in the sample composition and various unknowns about the longitudinal reliability of the test suggest some caution in interpreting these results.

Extrapolation

An extrapolation inference allows considering the score as a predictor of performance in the future or in another domain. This is of fundamental importance for the CAE because, at its core, the purpose of the CLA+ is to measure student readiness for employment by focusing on skills that are deemed to be required in the workplace (Benjamin et al., 2013).

Predicting performance

Supporting evidence

Two studies could be retrieved linking CLA/CLA+ scores to outcomes in the job market. Arum, Cho, Kim, and Roksa (2012) surveyed the enrolment, employment status and income of 925 new graduates who had taken the CLA in their first and final year. Among other findings, the authors reported that students in the bottom quintile of the CLA 'were three times more likely to be unemployed […] than those who performed in the top quintile' (p. 7), were 'significantly' more likely to have credit card debts (p. 3) and twice as likely to still live at home.

In a similar study, Zahner and James (2015) surveyed over 1,500 recent graduates, and found that CLA+ scores and race correlated with employment and postgraduate participation. The authors interpreted these results as evidence of the CLA+ predictive validity as well as of the existence of 'racial biases with respect to hiring, salary and enrolment in continuing education' (Zahner & James, 2015, p. 2).

Threats to validity

Arum et al. (2012) did not claim or imply that differences in post-university outcomes for students in the top and bottom CLA quintiles might be *caused* by different levels of CT skills, and with good reasons.

At an earlier stage of the study, they had shown that CLA scores correlated with variables such as student ethnicity, SAT scores, as well as with the performativity and segregation level of secondary schools (Arum & Roksa, 2011). These factors are interrelated. For example, 59% of Black students and 36% of Hispanic students were in the bottom SAT quintile, against 9% White students (Arum & Roksa, 2011, Table A2.2); because of this, 66% of black students attended less selective HEIs (Table A2.4).

While the follow-up report does not provide a breakdown of CLA performance by institution and demographic characteristics, it shows that less selective institutions had a higher unemployment rate than highly selective ones. Black and Hispanic students were also more likely to have taken college loans, live at home and have credit card debt (Arum et al., 2012), much like students in the bottom CLA quintile.

One way to read these data is that those who had been at a disadvantage while in education were both at greater risk of unemployment and happened to be in the lowest CLA quintile for reasons not necessarily linked with CT.

Racial biases were confirmed by Zahner and James (2015). The authors also maintained that the correlation CLA+ score – employment was evidence of the predictive power of the CLA+, but their article did not include information on the modelling approach, regression coefficients, errors or confidence levels. It would have been interesting to compare the effect sizes of the race and CLA+ coefficients; or to analyse whether the strength of the correlation between CLA+ and post-university outcomes would be reduced if one controlled for grades or SAT scores, which were not part of the model.

Taken together, findings from these studies confirm the existence of structural inequalities in the US. They also show that students who deal with tougher life contexts do less well both in the assessment and in higher education generally. This may be because doing well both in the CLA+ and in many university examinations entails being able to read long texts, to write argumentative essays and to review documentary evidence. In other words, while it is possible that CLA+ scores might be a good proxy for general academic ability,[3] the documents reviewed fall short of providing convincing evidence they are a good proxy of CT competence applied to a range of domains.

Decision-making

The *decision inference* regulates a social contract with test-takers: if a candidate's score means that the candidate has certain skills and knowledge, society grants him or her benefits such as access to education that was previously out-of-bounds, or it initiates remedial action in

the case skills are deemed to be unacceptably low. The warrant is that assessment purposes are clearly explained, under the assumption that uses will be in line with the stated purposes.

Test purposes and uses

Supporting evidence

In general, the CAE is very attentive to the practical applications of the CLA+. More than one section of the website is dedicated to explaining to a range of potential users the opportunities that the CLA+ offers. The 'trademark goal' of the CLA+ is to provide HEIs with a measure of value added growth, both at an aggregate level and at the level of the individual student (Benjamin et al., 2013, p. 2; also CAE, 2017; Zahner, 2014a).

Other purposes/uses are also suggested, including:

- To diagnose student deficits in CT skills.
- To benchmark initial and final performance in CT.
- To compare individual students, groups of students or institutions.
- To certify student CT proficiency or achievement.
- To inform curricular design.
- To evaluate the efficacy of undergraduate courses.
- To demonstrate faculty or university quality for accreditation or accountability.

For students, to provide employers with evidence of CT/work-readiness competence.

Threats to validity

There are several issues with the decision inference and its warrant, mainly linked to the claim that the CLA+ is a versatile instrument well-suited for a wide range of uses. Newton (2007) warned about the validity of one-size-fits-all assessments, since different purposes require different designs. Indeed, because of the commercial nature of the CLA+, many claims about its qualities are simply promotional statements that cannot stand to proper scrutiny.

For example, the CLA+ does use 'proficiency standard levels defined by experts from business, K-12, and higher education' (CAE, 2017); however, the experts consulted to define such standard levels were only 12 (Zahner, 2014b), which is hardly a representative sample. Likewise, it is fair to say that 'early detection of critical-thinking deficits helps individuals and institutions target further instruction' (CAE, 2017), but the information returned by the CAE in post-test reports is too little for an accurate diagnosis (see e.g. CAE, 2015b). Item-level information is unavailable because the items are copyrighted and secured, but student PT responses are also unavailable, even though they could be considered the students' intellectual property. Without access to student responses to individual items (or to the items themselves), it is not possible to provide tailored formative feedback.

A conflict between advertisement and what the CLA+ can *really* be used for emerges in the institutional report (CAE, 2015b). After noting that 'CLA+ results provide a valuable tool for potential employers and graduate schools to ascertain the depth of a student's critical-thinking and written-communication skills' (p. i); and that 'educators may decide to consult their students' CLA+ results when making individualized decisions related to admission, placement, scholarships, or grading' (p. 7); the CAE adds the following contradictory disclaimer:

> Institutions should not use mastery levels for purposes other than the interpretation of test results. If an institution wishes to use the attainment of CLA+ mastery levels as part of a graduation requirement or the basis for an employment decision, the institution should conduct a separate standard-setting study with this specific purpose in mind. (CAE, 2015b, p. 16)

Besides, Steedle (2012) had already dismissed the possibility of using the CLA+ for high-stakes decisions since 'reliability around 0.75 is not likely adequate' (p. 649). Year-on-year consistency is also currently not acceptable for this use.

Even decisions based on the CLA's original purpose, making cross-institutional comparisons, must be carefully considered. Comparisons are based on the value added by participating institutions. Technically, these 'value-added' scores are the standardised level-2 residuals from a multilevel model regressing finalist scores on finalist entry achievement and on first-year students' CLA+ scores.

This approach finds ample use in educational literature (Goldstein, 2011), but it has limitations when informing policy decisions. This is because level-2 residuals are assumed to be normally distributed. If the assumption holds, then every year about 32% of all institutions would fall beyond ±1 standard deviations from the mean. In practice, institutions could set relative targets such as 'let us try to have a positive residual next year, provided the participating institutions remain the same', but higher-level governmental objectives like '50% of universities should exceed expectations' are not achievable.

Discussion

The CLA+ has many positive features. Eighteen students at Reading responded to a questionnaire to gauge their opinion of the test. Although few students completed this, a number did comment that the PT was both interesting and challenging. The combination of a problematic scenario with a documentary review requires the application of a complex network of specialised skills that may be highly valued in some work contexts. The online administration worked smoothly, and the team at CAE was supportive, passionate and reachable.

Nevertheless, some threats to the validity of the CLA+ were identified. This article focused on actionable issues, threats these authors think the CAE could address and that might have a negative impact on learning gain in the UK.

The first is the very definition of CT, for which there is no consensus. Therefore, claiming alignment with three (not completely compatible) definitions is unsatisfactory, particularly when the assessment appears to have developed from more solid grounds. Most of the times, the CAE treats what is credited by the scoring rubric as if it were the construct, whereas it should demonstrate why it believes that measuring certain aspects of student performance *entails* measuring the construct. The test can capture academic abilities that are useful in liberal arts (as noted by Arum et al., 2011) and that correlate with student socio-economic and demographic characteristics. Whether these abilities can be called 'critical thinking' skills useful to measure relative learning gain in UK HEIs is still unclear.

A second issue concerns some technical aspects of the assessment. The defining feature of the CLA+ is the PT, which attempts to simulate a plausible workplace scenario and has high 'structural fidelity' (Kuechler & Simkin, 2010). However, this holistic approach is at odds with the analytic format of the scoring rubric and with the summing and averaging criteria underpinning the scoring system. Good writing skills may inflate the total score, creating a mismatch between observed performance and inferences regarding a student's

ability. In practice, this means that it is usually more informative to consider the PT and the SRQ as two separate assessments. The test is internally consistent, but student-level reliability is still too low to warrant its use for student-level decisions. Some type of value-added measures (level-2 residuals) are of limited use for policy-making. It is also unclear which of the suggested uses are promotional statements, and which can follow from the interpretation of test scores.

From a practical angle, there is a question about the role the CLA+ might serve as part of a suite of measures of learning gain in a UK university. On an institutional level, CLA+ scores correlate with other measures of academic achievement; some of them are more readily available, cheaper to obtain and do not involve the administrative burden of student recruitment and testing. On an individual level, the longitudinal reliability is too low to detect changes in student performance consistently, which would invalidate using the test to measure learning gain. One could administer the CLA+ several times and look at the overall trend, but this would be costly.

Extending the critique beyond the CLA+, it is worth considering whether CT skills are the right measure for learning gain in higher education. From an accountability perspective, the extent to which HEIs would be able to affect student competence is unclear. Systematic reviews of the efficacy of teaching methods to improve CT in nursing and social education found mixed evidence and noted the varying quality of the few studies reviewed (Brudvig, Dirkes, Dutta, & Rane, 2013; Carter, Creedy, & Sidebotham, 2016; Kong, Qin, Zhou, Mou, & Gao, 2014; Lee, Lee, Gong, Bae, & Choi, 2016; Samson, 2016). A meta-analysis in the US suggested that CT skills increase on average by 0.59 standard deviations after four years of higher education, but courses in which critical thinking are explicitly taught in the curriculum (e.g. nursing) did no better than the rest (similar findings were reported in Brudvig et al., 2013; and in Niu, Behar-Horenstein, & Garvan, 2013).

One should therefore consider whether the CT paradigm the CLA+ and many US tests subscribe to – sometimes called 'logicistic', because CT is viewed as an application of informal logic[4] (Walker & Finney, 1999) – is in line with the deeper purposes of higher education. Critical thinking is not just about 'evaluating the credibility of texts or in problem solving, but mainly involves critical analysis of social, economic, and political implications of texts to promote a more just world' (Ibrahim, 2015, p. 756).

One of the selling features of the CLA+ is its purported ability to measure skills that are necessary for country economy and are highly valued by employers. A survey piloted in nine European countries showed that employers favour skills such as reading literacy, team working and the ability to respond to instruction (Cedefop, 2013), though. What employers want is often highly situated and nuanced, and cannot be addressed by catch-all statements (Frankham, 2016) and therefore alignment of CLA+ scores with employability and learning gain is not robust.

Conclusions

Learning gain has become a core part of the Government's plans for higher education (HEFCE, 2016) and is identified as one of the three major categories in the Year 2 TEF, along with 'Teaching Quality' and the 'Learning Environment'. It has developed behind issues in the US over the value of the time and financial investment in higher education. With the advent of knowledge economy, CT skills have been framed as *the* missing link between

higher education and employment and as such robust measurement of these skills would be a good proxy for learning gain. The CLA+ is one of several products competing in the market to offer a measurement of CT, and this article is a first attempt to highlight some of its possible shortcomings.

Of course, a full validation might reveal that these threats are not as serious as to undermine the overall validity of the assessment, and these authors would encourage further research on this topic. For the moment, however, there seems to be a lack of balance between the administrative and financial commitment of delivering the test to a wide student population twice a year (to entering and exiting students), and the extent to which it can offer the required information for learning gain purposes.

Notes

1. Broadly speaking, an argument is valid if the internal logic is consistent and the conclusion follows from its premises, but it is sound if it is valid and the premises are true (Roy, 2017).
2. It is unclear why Zahner and Steedle (2015) decided to use two different model equations. The authors seem convinced that the two models are substantively different; so much that they call the first 'CLA value-added' model and the second a 'random effects' model. In fact, both are random effects models. The level-2 residuals are shrunken in both cases, whereas Zahner and Steedle (2015) seem to suggest that they are not. The two models differ only in the choice of covariates; when the authors switch datasets, the only difference between the models is that one does not include the aggregate SATs scores. It is therefore unsurprising to observe high correlations between predictions.
3. This relationship would not be unique to the CLA+. For example, the California Critical Thinking Skills Test was found to be correlated to university marks (O'Hare & McGuinness, 2015).
4. Notice that even logicists do not consider CT *only* as a matter for logic. The ethics of critical thinking were discussed in Facione (1990), as was its being a 'liberating' force (p. 2). There is wide agreement that being rational 'also requires an open-minded yet critical approach to one's own thinking as well as that of others' (Black, 2012, p. 125), and critical thinking skills have been viewed as playing an important role in 'solid liberal education' (Facione, 1990, p. 5). Bailin and Siegel (2003) suggested that 'having the ability to think critically requires […] having the ability to ascertain the goodness of candidate reasons [for or against a judgement]' (p. 182), but they also acknowledged that the criteria whereby a reason is to be considered good need not draw exclusively from the sphere of logic.

Funding

This study was part of the Learning Gain project, funded by the Higher Education Funding Council for England (HEFCE) [grant number 10007802].

ORCID

Cesare Aloisi http://orcid.org/0000-0002-7151-7379
A. Callaghan http://orcid.org/0000-0002-2731-3352

References

ACT. (2017). *ACT collegiate assessment of academic proficiency* [Web page]. Retrieved from http://www.act.org/content/act/en/products-and-services/act-collegiate-assessment-of-academic-proficiency.html

Alexander, P.A. (2014). Thinking critically and analytically about critical-analytic thinking: An introduction. *Educational Psychology Review, 26*(4), 469–476. doi:10.1007/s10648-014-9283-1

Arum, R. (2013). Stakeholder and public responses to measuring student learning. *Society, 50*(3), 230–235. doi:10.1007/s12115-013-9648-y

Arum, R., Cho, E., Kim, J., & Roksa, J. (2012). *Documenting uncertain times: Post-graduate transitions of the Academically Adrift cohort*. New York, NY: Social Science Research Council.

Arum, R., & Roksa, J. (2011). *Academically Adrift: Limited learning on college campuses*. Chicago, IL: The University of Chicago Press.

Arum, R., Roksa, J., & Cho, E. (2011). *Improving undergraduate learning: Findings and policy recommendations from the SSRC-CLA longitudinal project*. New York, NY: Social Science Research Council.

Bailin, S., & Siegel, H. (2003). Critical thinking. In N. Blake, P. Smeyers, R. Smith, & P. Standish (Eds.), *The Blackwell guide to the philosophy of education* (pp. 181–193). Oxford: Blackwell.

Barbaranelli, C., Lee, C.S., Vellone, E., & Riegel, B. (2015). The problem with Cronbach's Alpha: comment on Sijtsma and van der Ark (2015). *Nursing Research, 64*(2), 140–145. doi:10.1097/NNR.0000000000000079

Bassok, M., & Novick, L.R. (2012). Problem solving. In K.J. Holyoak & R.G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 1–22). Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199734689.013.0021

Bates, D.M. (2007, November 10). *Re: [R] Confidence intervals for random effect BLUP's*. [Electronic mailing list message]. Retrieved from http://www.mail-archive.com/r-help@r-project.org/msg04913.html

Bates, D.M. (2010). *lme4: Mixed-effects modeling with R*. Retrieved from http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf

Benjamin, R., Klein, S.P., Steedle, J.T., Zahner, D., Elliot, S., & Patterson, J. (2013). *The case for critical-thinking skills and performance assessment*. New York, NY: Council for Aid to Education.

Berkow, S., Virkstis, K., Stewart, J., Aronson, S., & Donohue, M. (2011). Assessing individual frontline nurse critical thinking. *The Journal of Nursing Administration, 41*, 168–171.

Black, B. (2012). An overview of a programme of research to support the assessment of critical thinking. *Thinking Skills and Creativity, 7*(2), 122–133. doi:10.1016/j.tsc.2012.04.003

Bok, D.C. (2006). *Our underachieving colleges: A candid look at how much students learn and why they should be learning more*. Princeton, NJ: Princeton University Press.

Brudvig, T.J., Dirkes, A., Dutta, P., & Rane, K. (2013). Critical thinking skills in health care professional students: A systematic review. *Journal of Physical Therapy Education, 27*(3), 12–25.

Brunt, B.A. (2005). Models, measurement, and strategies in developing critical-thinking skills. *The Journal of Continuing Education in Nursing, 36*(6), 255–262.

Byrnes, J.P., & Dunbar, K.N. (2014). The nature and development of critical-analytic thinking. *Educational Psychology Review, 26*(4), 477–493. doi:10.1007/s10648-014-9284-0

Cacioppo, J.T., Petty, R.E., Feinstein, J.A., & Jarvis, W.B.G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119*(2), 197–253. doi:10.1037/0033-2909.119.2.197

Cambridge Assessment Admissions Testing. (2017). *Thinking skills assessment*. Retrieved from http://www.admissionstesting.org/for-test-takers/thinking-skills-assessment/

Cambridge International Examinations. (2016). *Frequently Asked Questions (FAQs): Cambridge International AS & A Level Thinking Skills (9694)*. Retrieved from http://www.cambridgeinternational.org/images/131019-frequently-asked-questions.pdf

Carter, A.G., Creedy, D.K., & Sidebotham, M. (2015). Evaluation of tools used to measure critical thinking development in nursing and midwifery undergraduate students: A systematic review. *Nurse Education Today, 35*(7), 864–874. doi:10.1016/j.nedt.2015.02.023

Carter, A.G., Creedy, D.K., & Sidebotham, M. (2016). Efficacy of teaching methods used to develop critical thinking in nursing and midwifery undergraduate students: A systematic review of the literature. *Nurse Education Today, 40*(7), 209–218. doi:10.1016/j.nedt.2016.03.010

Clark, L.A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309–319. doi:10.1037/1040-3590.7.3.309

Council for Aid to Education. (2014a). *CLA+: Technical FAQ*. New York, NY: Author.
Council for Aid to Education. (2014b). *The case for performance-based assessment of critical-thinking skills*. Retrieved from http://cae.org/flagship-assessments-cla-cwra/flagship-assessments-cla-cwra/resources/
Council for Aid to Education. (2015a). *About CAE: History* [Web page]. Retrieved from http://cae.org/about/history/
Council for Aid to Education. (2015b). *Spring 2015 CLA+ mastery results: Sample university (Institutional report)*. New York, NY: Author.
Council for Aid to Education. (2016). *CLA+: National results, 2015–16*. Retrieved from http://cae.org/flagship-assessments-cla-cwra/cla/resources-for-cla/
Council for Aid to Education. (2017). *Why CLA+ and CWRA+* [Web page]. Retrieved from http://cae.org/flagship-assessments-cla-cwra/why-cla-and-cwra/
Cedefop. (2013). *Piloting a European employer survey on skill needs: Illustrative findings* (Research paper No. 36). Luxembourg: Publications Office of the European Union. doi:10.2801/3701
College Board. (2017). *SAT suite of assessments* [Web page]. Retrieved from https://collegereadiness.collegeboard.org/sat
Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. doi:10.1037/0021-9010.78.1.98
Coutinho, S.A. (2006). The relationship between the need for cognition, metacognition, and intellectual task performance. *Educational Research and Reviews, 1*(5), 162–164.
Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. doi:10.1037/h0040957
Dalkey, N., & Helmer, O. (1962). *An experimental application of the Delphi method to the use of experts (Memorandum RM-727/1-abridged)*. Santa Monica, CA: RAND Corporation.
Del Bueno, D.J. (1990). Experience, education, and nurses' ability to make clinical judgments. *Nursing and Health Care, 11*, 290–294.
Elder, L. (2007). *Another brief conceptualization of critical thinking* [Web page]. Retrieved from http://www.criticalthinking.org/pages/defining-critical-thinking/766
Ennis, R.H. (1962). A concept of critical thinking. *Harvard Educational Review, 32*(1), 81–111.
Ennis, R.H., Millman, J., & Tomko, T.N. (1985). *Cornell critical thinking tests level X and level Z manual*. (3rd ed.). Pacific Grove, CA: Midwest Publications.
Ennis, R.H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. Baron & R. Sternberg (Eds.), *Teaching thinking skills theory and practice* (pp. 9–26). New York, NY: W. H. Freeman.
Ennis, R.H., & Weir, E. (1985). *The Ennis-Weir critical thinking essay test*. Pacific Grove, CA: Midwest Publications.
Erwin, T.D., & Sebrell, K.W. (2003). Assement of critical thinking: ETS's tasks in critical thinking. *The Journal of General Education, 52*(1), 50–70.
ETS. (2017). *ETS proficiency profile* [Web page]. Retrieved from https://www.ets.org/proficiencyprofile/about/
Evans, J.S.B.T., & Stanovich, K.E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241. doi:10.1177/1745691612460685
Facione, P.A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction – The Delphi Report* (Executive Summary). Millbrae, CA: The California Academic Press.
Facione, P.A. (1992). *The California Critical Thinking Skills Test manual*. Millbrae, CA: California Academic Press.
Facione, P.A., & Facione, N.C. (1992). *The California Critical Thinking Dispositions Inventory test manual*. Millbrae, CA: California Academic Press.
Frankham, J. (2016). Employability and higher education: The follies of the "productivity challenge" in the Teaching Excellence Framework. *Journal of Education Policy,* 1–14. doi:10.1080/02680939.2016.1268271

Furnham, A., & Thorne, J.D. (2013). Need for cognition: Its dimensionality and personality and intelligence correlates. *Journal of Individual Differences, 34*(4), 230–240. doi:10.1027/1614-0001/a000119

Glenn, D. (2011, January 18). New book lays failure to learn on colleges' doorsteps. *The Chronicle of Higher Education*. Retrieved from http://www.chronicle.com/article/New-Book-Lays-Failure-to-Learn/125983/

Goldstein, H. (2011). *Multilevel statistical models*. (4th ed.). Chichester: Wiley.

Halpern, D.F. (2010). Halpern Critical Thinking Assessment manual. Vienna: Schuhfried GmbH.

Halpern, D.F. (2013). *Critical thinking across the curriculum: A brief edition of thought and knowledge*. New York, NY: Routledge.

Higher Education Funding Council for England. (2016). *Learning and teaching excellence: Learning gain* [Web page]. Retrieved from http://www.hefce.ac.uk/lt/lg/

Higher Education Funding Council for England. (2017). *Qualifications obtained* [Web page]. Retrieved from https://www.hesa.ac.uk/data-and-analysis/students/qualifications

Ibrahim, N.K. (2015). Critical literacy: Performance and reactions. *Theory and Practice in Language Studies, 5*(4), 756–764. doi:10.17507/tpls.0504.11

Insight Assessment. (2013). *California critical thinking disposition inventory (CCTDI)*. San Jose, CA: The Californian Academic Press.

Johnson, R.H., & Hamby, B. (2015). A meta-level approach to the problem of defining "critical thinking". *Argumentation, 29*(4), 417–430. doi:10.1007/s10503-015-9356-4

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. doi:10.1111/jedm.12000

Klein, S.P. (1996). The costs and benefits of performance testing on the bar examination. *The Bar Examiner, 65*(3), 13–20.

Klein, S.P. (2008). Characteristics of hand and machine-assigned scores to college students' answers to open-ended tasks. In *Probability and statistics: Essays in honor of David A. Freedman* (Vol. 2, pp. 76–89). Beachwood, OH: Institute of Mathematical Statistics. doi:10.1214/193940307000000392

Klein, S.P., Benjamin, R., Shavelson, R.J., & Bolus, R. (2007). The Collegiate Learning Assessment: Facts and fantasies. *Evaluation Review, 31*(5), 415–439. doi:10.1177/0193841X07303318

Klein, S.P., Kuh, G.D., Chun, M., Hamilton, L., & Shavelson, R. (2005). An approach to measuring cognitive outcomes across higher education institutions. *Research in Higher Education, 46*(3), 251–276. doi:10.1007/s11162-004-1640-3

Klein, S.P., Liu, O.L., & Sconing, J. (2009). *Test Validity Study (TVS) report*. Retrieved from http://cae.org/images/uploads/pdf/13_Test_Validity_Study_Report.pdf

Kong, L.-N., Qin, B., Zhou, Y., Mou, S., & Gao, H.-M. (2014). The effectiveness of problem-based learning on development of nursing students' critical thinking: A systematic review and meta-analysis. *International Journal of Nursing Studies, 51*(3), 458–469. doi:10.1016/j.ijnurstu.2013.06.009

Kuechler, W.L., & Simkin, M.G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education, 8*(1), 55–73. doi:10.1111/j.1540-4609.2009.00243.x

Lai, E.R. (2011). *Critical thinking: A literature review*. Retrieved from http://images.pearsonassessments.com/images/tmrs/CriticalThinkingReviewFINAL.pdf

Lawson, T.J., Jordan-Fleming, M.K., & Bodle, J.H. (2015). Measuring psychological critical thinking. *Teaching of Psychology, 42*(3), 248–253. doi:10.1177/0098628315587624

Lee, J., Lee, Y., Gong, S., Bae, J., & Choi, M. (2016). A meta-analysis of the effects of non-traditional teaching methods on the critical thinking abilities of nursing students. *BMC Medical Education, 16*(1), 240. doi:10.1186/s12909-016-0761-7

Lindsay, T.K. (2013). *The likelihood of higher-education reform. Society, 50*(3), 236–244. doi:10.1007/s12115-013-9649-x

Lipman, M. (1987). Critical thinking: What can it be? *Analytic Teaching, 8*(1), 5–12.

Liu, O.L., Frankel, L., & Roohr, K.C. (2014). Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series, 2014*(1), 1–23. doi:10.1002/ets2.12009

Liu, O.L., Mao, L., Frankel, L., & Xu, J. (2016). Assessing critical thinking in higher education: The HEIghtenTM approach and preliminary validity evidence. *Assessment & Evaluation in Higher Education, 41*(5), 677–694. doi:10.1080/02602938.2016.1168358

McGrath, C.H., Guerin, B., Harte, E., Frearson, M., & Manville, C. (2015). *Learning gain in higher education*. Santa Monica, CA: RAND Corporation.

McKelvie, S.J. (1992). Does memory contaminate test – Retest reliability? *Journal of General Psychology, 119*(1), 59–72.

McPeck, J.E. (1981). *Critical thinking and education*. New York, NY: St. Martin's Press.

Menchaca, F. (2014). Start a new fire: Measuring the value of academic libraries in undergraduate learning. *Portal: Libraries and the Academy, 14*(3), 353–367. doi:10.1353/pla.2014.0020

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*(3), 215–237.

Messick, S. (1989). Validity. In: R.L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York, NY: Macmillan.

Murphy, P.K., Rowe, M.L., Ramani, G., & Silverman, R. (2014). Promoting critical-analytic thinking in children and adolescents at home and in school. *Educational Psychology Review, 26*(4), 561–578. doi:10.1007/s10648-014-9281-3

Newton, P.E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education, 14*(2), 149–170. doi:10.1080/09695940701478321

Newton, P.E., & Shaw, S.D. (2015). Disagreement over the best way to use the word "validity" and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice, 23*(2), 178–197. doi:10.1080/0969594X.2015.1037241

Niu, L., Behar-Horenstein, L.S., & Garvan, C.W. (2013). Do instructional interventions influence college students' critical thinking skills? A meta-analysis. *Educational Research Review, 9*, 114–128. doi:10.1016/j.edurev.2012.12.002

O'Hare, L., & McGuinness, C. (2015). The validity of critical thinking tests for predicting degree performance: A longitudinal study. *International Journal of Educational Research, 72*, 162–172. doi:10.1016/j.ijer.2015.06.004

Opposs, D., & He, Q. (2011). *The reliability programme: Final report of the technical advisory group* (No. 4828). Coventry: Office of Qualifications and Examinations Regulation.

Pascarella, E.T., & Terenzini, P.T. (2005). *How college affects students: a third decade of research*. San Francisco, CA: John Wiley and Sons.

Paul, R. (1992). Critical thinking: What, why, and how. *New Directions for Community Colleges, 1992*(77), 3–24. doi:10.1002/cc.36819927703

Popham, W.J. (1997). What's wrong – And what's right – With rubrics. *Educational Leadership, 55*(2), 72–75.

Pucer, P., Trobec, I., & Žvanut, B. (2014). An information communication technology based approach for the acquisition of critical thinking skills. *Nurse Education Today, 34*(6), 964–970.

Romeo, E.M. (2010). Quantitative research on critical thinking and predicting nursing students' NCLEX-RN performance. *Journal of Nursing Education, 49*(7), 378–386. doi:10.3928/01484834-20100331-05

Roy, T. (2017). *Symbolic logic: A accessible introduction to serious mathematical logic*. (Version 7.3). Retrieved from http://rocket.csusb.edu/~troy/SLmain.pdf

Sadler, D.R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159–179. doi:10.1080/02602930801956059

Samson, P.L. (2016). Critical thinking in social work education: A research synthesis. *Journal of Social Work Education, 52*(2), 147–156. doi:10.1080/10437797.2016.1151270

Scriven, M., & Paul, R. (1987). *Critical thinking. 8th Annual International Conference on Critical Thinking and Education Reform* [Web page]. Retrieved from http://www.criticalthinking.org/pages/defining-critical-thinking/766

Shavelson, R.J., & Huang, L. (2003). Responding responsibly to the frenzy to assess learning in higher education. *Change, 1383*(January/February), 11–19. doi:10.1080/00091380309604739

Shaw, S.D., & Crisp, V. (2012). An approach to validation: Developing and applying an approach for the validation of general qualifications. *Research Matters: A Cambridge Assessment Publication*, Special

Issue 3. Retrieved from http://www.cambridgeassessment.org.uk/our-research/all-published-resources/research-matters/

Siegel, H. (1988). *Educating reason: Rationality, critical thinking and education*. New York, NY: Routledge.

Sireci, S.G. (1998). The construct of content validity. *Social Indicators Research, 45*(1), 83–117. doi:10.1023/A:1006985528729

Sosu, E.M. (2013). The development and psychometric validation of a Critical Thinking Disposition Scale. *Thinking Skills and Creativity, 9*, 107–119. doi:10.1016/j.tsc.2012.09.002

Stanovich, K.E. (2016). The comprehensive assessment of rational thinking. *Educational Psychologist, 51*(1), 23–34. doi:10.1080/00461520.2015.1125787

Stedman, N.L.P., Irani, T.A., Friedel, C., Rhoades, E.B., & Ricketts, J.C. (2009). Relationships between critical thinking disposition and need for cognition among undergraduate students enrolled in leadership courses. *NACTA Journal, 53*(3), 62–70.

Steedle, J.T. (2012). Selecting value-added models for postsecondary institutional assessment. *Assessment & Evaluation in Higher Education, 37*(6), 637–652. doi:10.1080/02602938.2011.560720

Steedle, J.T. (2014). Motivation filtering on a multi-institution assessment of general college outcomes. *Applied Measurement in Education, 27*(1), 58–76. doi:10.1080/08957347.2013.853072

Stupple, E.J.N., Maratos, F.A., Elander, J., Hunt, T.E., Cheung, K.Y.F., & Aubeeluck, A.V. (2017). Development of the Critical Thinking Toolkit (CriTT): A measure of student attitudes and beliefs about critical thinking. *Thinking Skills and Creativity, 23*, 91–100. doi:10.1016/j.tsc.2016.11.007

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. doi:10.5116/ijme.4dfb.8dfd

Thayer-Bacon, B. (1998). Transforming and redescribing critical thinking: Constructive thinking. *Studies in Philosophy and Education, 17*(2/3), 123–148. doi:10.1023/A:1005166416808

Walker, P., & Finney, N. (1999). Skill development and critical thinking in higher education. *Teaching in Higher Education, 4*(4), 531–547. doi:10.1080/1356251990040409

Watson, G., & Glaser, E.M. (1980). *Watson-Glaser critical thinking appraisal manual*. Cleveland, OH: Psychological Corp.

Winterbotham, M., Vivian, D., Shury, J., Davies, B., & Kik, G. (2014). *UK commission's employer skills survey 2013: UK results* (Evidence report No. 81). Retrieved from https://www.gov.uk/government/publications/ukces-employer-skills-survey-2013

Wolf, R., Zahner, D., & Benjamin, R. (2015). Methodological challenges in international comparative post-secondary assessment programs: Lessons learned and the road ahead. *Studies in Higher Education, 40*(3), 471–481. doi:10.1080/03075079.2015.1004239

Zahner, D. (2014a). *Reliability and validity of CLA+*. New York, NY: Council for Aid to Education.

Zahner, D. (2014b). *Standard-setting study final report*. New York, NY: Council for Aid to Education.

Zahner, D., & James, J.K. (2015). *Predictive validity of a critical thinking assessment for post-college outcomes*. New York, NY: Council for Aid to Education.

Zahner, D., & Steedle, J.T. (2015). *Comparing longitudinal and cross-sectional school effect estimates in postsecondary education*. Paper presented at the 2015 National Council on Measurement in Education Annual Meeting, Chicago, IL.

Zuriguel Pérez, E., Lluch Canut, M.T., Falcó Pegueroles, A., Puig Llobet, M., Moreno Arroyo, C., & Roldán Merino, J. (2015). Critical thinking in nursing: Scoping review of the literature. *International Journal of Nursing Practice, 21*(6), 820–830. doi:10.1111/ijn.12347